A Conceptual Model for Clinical Radiology Reports

Carol Friedman Ph.D.[†]
James J. Cimino, M.D., Stephen B. Johnson, Ph.D.[‡]

[†] Queens College of the City University of New York
[‡] Columbia University, New York

Abstract

The structural and informational content of clinical radiology reports was examined to develop a comprehensive representational schema of the concepts in the domain. The model involves several different conceptual levels, ranging from the high level description of the report to the lower level description of the clinical concepts contained in the reports and the specification of the terms used to express the concepts. The design of an adequate structured representation for the domain has important implications for the design of the electronic patient record, for the unification of different controlled vocabularies by enabling them to be mapped to one common representation, and for the facilitation of natural language processing of clinical reports so that coded data may be obtained.

1 Introduction

We seek to accurately represent the clinical information in radiology reports in structured coded form for subsequent use in computer-assisted analyses, such as automated decision support, case coding and clinical research. This paper briefly describes components of the model we have developed for this task. One of our goals is to develop a representational model which supports natural language processing so that the clinical information in the radiology reports may be automatically structured and encoded using natural language methodology. However in this paper we focus on the representation issues and not on the language processing methodology. Another goal is to develop a model which moves toward a canonical representation of medical terminology so that medical terms from one controlled vocabulary can be mapped into the structured representation of medical concepts in our model. The model we are proposing has been designed by analyzing the structure and semantic contents of Chest Xray Reports (CXRs) obtained from the central textual patient database at Columbia Presybterian Medical Center (CPMC).

Other articles that discuss the management of medical terminology ([6, 10, 8, 7, 1, 2]) have focussed on important issues such as representation of medical concepts, taxonomies, knowledge, and meta-knowledge. In this paper we focus on a structured and compositional representation of medical concepts. When concepts are identified by multi-word terms they usually consist of combinations of simpler concepts. Using our model it is possible to identify complex concepts from their simpler components.

2 Background

The structure of information present in radiology reports can be viewed at four different conceptual levels. The first level is the representation of the structure of the report itself. This is described in more detail in Section 3. The second level is the representation of the findings in the report, which have their own complex structure of medical concepts. A finding in a report may not necessarily correspond solely to one medical concept, but may correspond to one concept that is associated with modifier concepts. For example, *mild cardiomegaly* consists of the finding *cardiomegaly* modified by a severity modifier which has the value *mild*.

The third level is the structure of the medical concepts that make up the findings. Some of these concepts may be basic lexical concepts that consist of one word, such as *infiltrate*. However, other concepts consist of combinations of several more basic medical concepts. These combinations do not always co-occur in exactly the same order when they appear in the text. For example, if we were to assume that the term *worsening left pleural effusion* formed a unique concept in our vocabulary, it would consist of the finding concept *pleural effusion* modified by a temporal concept *worse*, and a laterality concept *left*. Our model would represent the compositionality of the concept *worsening left pleural effusion* in a structured representation so that if that same concept (possibly occurring with additional modifiers) were expressed slightly differently, as in *left pleural effusion appears worse*, it would be possible to identify the correct concept from the components. By structuring the expression and by matching its structured form against the structured forms of the concepts in our controlled vocabulary, the concept *worsening left pleural effusion* would be found to have a structure which is contained in the structure of *left pleural effusion appears worse*. The latter expression would therefore be mapped to the concept *worsening left pleural effusion*,

0195-4210/92/$5.00

with a certainty modifier corresponding to *appears*.

The conceptual representations describing the report structure, the finding structure, and the medical concept structure are all represented within the framework of the Medical Entities Dictionary (MED) [6] developed at CPMC. The MED contains unique concepts that form a semantic network which supports multiple inheritance. Each concept is represented as a frame with slots which specify predefined relations to the concept. Below, we use the linear notation for Conceptual Graphs [9] because a frame can be precisely represented as a conceptual graph (CG) (see Appendix A. for a description of CG notation).

The fourth conceptual level is the lexical information associated with individual words and multi-word phrases used in the reports. The lexicon classifies terms according to their semantic classes, but the same classes must also be present in the MED. The lexicon also specifies regularized forms for terms. For example, the semantic class of both *enlarged* and *enlargement* is **Descriptor**, and the regularized form for each is **enlarged**. Similarly the semantic class for both *heart* and *cardiac* is **Bodyloc**, and the regularized form for each is **heart**. Multi-word terms are also found in the lexicon, particularly if they occur frequently in the domain. The multi-word term *elevation of diaphragm* has the semantic category **Rad Finding**.

Because terms in the lexicon do not necessarily correspond to concepts in the MED, we need a component of the model which forms a bridge between the terms as expressed in the text and the concepts in the MED. This is accomplished by means of a synonym table. Although each concept in the MED can specify a list of synonyms, for convenience we maintain a separate synonym table. If the regularized form of a term is not in the MED, a synonym is provided for the term that is equivalent to a concept in the MED. For example, in the lexicon, **dyspnea** is the regularized form for the word *dyspnea*, and the corresponding MED concept is **dyspnea**. Therefore no entry for **dyspnea** is required in the synonym table. However, there is no matching MED concept for **shortness of breath** which is the regularized form for the terms *shortness of breath* and *breathlessness*. In this case, there would be an entry in the synonym table for **shortness of breath** specifying that **dyspnea** is the corresponding MED concept.

3 The Report Structure

Manual analysis of the radiology reports revealed that they have a structure which places the concepts in a variety of contexts. In these contexts, the concepts retain their meanings but their implications for uses such as decision support may vary. For example, a report may indicate information given to the radiologist about the patient, as well as observations made in the description of the film, and interpretations based on the description. Findings can appear in any of these contexts and therefore a single model is used to represent them, while retaining the nature of the context. Concepts may also appear in radiology reports in other contexts, as reasons for the exam, information about the patient, the date of the report, etc. In order to simplify the model in this paper, these types of information have been omitted below. A simplified version of the description of a CXR structure is shown below:

```
[Chest Xray Report] -
  (Proc Type)->[Chest Xray Proc:@1]
  (Proc Location)->[chest:@1]
  (Comp Report)->[Chest Xray Report:{*}]
  (Description)->[Xray Finding Sent:{*}]
  (Impression)->[Xray Finding Sent:{*}].
```

The **Proc Type** relation represents the chest xray procedure that was performed, the **Proc Location** relation has the concept [chest] as its value, and the **Comp Report** relation provides for the inclusion of a reference to a previous report if such appears in the current report. The final two slots, **Description** and **Impression**, correspond to the two sections of radiology reports which typically contain sentences consisting of findings. The description of [Xray Finding Sent] shown below consists of the original text of the report, as well as the structured finding(s). In our system, the structured finding is the output produced by the natural language processor from the text of the report. However, structured findings may also be produced by other means, such as structured data entry. Our definition of **Xray Finding Sent** is:

```
[Xray Finding Sent]-
  (Text)->[String Data:@1]
  (Structured)->[Rad Finding Struct:@>0].
```

4 The Findings Structure

Analysis of radiology findings showed them to be complex arrangements of basic medical concepts. The possible permutations of radiology findings suggest that enumerating them would be impractical. However, the relations between concepts in each radiology finding appear to be of a relatively small number, and the clinical content of radiology findings may be adequately structured. The structure **Rad Finding Struct**, which consists of a central finding concept with optional modifiers, is shown below:

```
[Rad Finding Struct]-
  (Central Finding)->[Rad Finding Struct:{*}]
  (Bodyloc Mod)->[Bodyloc:{*}]
  (Finding Mod)->[Modifier:{*}]
  (Related Finding)->[Relational Finding:{*}]
  (Evidential Proc)->[Surgical Proc:{*}]
  (Technique Info)->[Technique:{*}]
  (Management Info)->[Management Proc:{*}].
```

Radiology findings interact with each other in very limited ways. A reference to another radiology finding is one way in which the current radiology finding may express a modification of a finding noted in a previous radiology report. The information that the finding was

noted in a previous report is represented by the Finding Mod relation which has a modifier which is a temporal type. The information may say that the finding was previously noted, as in *markings were previously noted.* In this case, the value of temporal modifier would be **previous**, denoting that the concept **markings** occurred in a previous report. The temporal information may represent a change *increased* in a finding, as in *opacity has increased.*

Another way in which radiology findings may be related is when one radiology finding suggests a second finding (as in *markings are consistent with atelectasis*). Since the second finding is not a direct observation in the report, but rather included as related to the first finding, the second finding is represented as part of the description of the first. However findings which are parallel (as in *markings and opacity noted, markings as well as opacity noted, markings with opacity*) are represented as multiple findings on the same level. The descriptions of the relations in **Rad Finding Struct** are as follows: a. **Central Finding**. Represents a concept that is the central part of the radiology finding. This may be a complete radiology finding concept **cardiomegaly**, or a descriptive concept **enlarged**, which is a partial finding which together with a **Bodyloc Mod** forms a complete radiology finding. A partial finding may be stored in this slot as a result of text processing which structures textual sentences, such as *heart appears enlarged.* If the value of **Central Finding** is not a complete radiology finding, by searching the compositional models for the complex concepts that are finding concepts, it may be possible to map the value **enlarged** into a complete radiology finding. This is explained in detail in Section 5. b. **Bodyloc Mod**. Represents the body location of the radiology finding. Thus, if the sentence is *heart appears enlarged*, the Bodyloc Mod would be **heart**. c. **Finding Mod**. In addition to finding concepts, radiology findings in a report typically have modifiers containing information about severity, certainty, temporal information, and quantitative concepts. For example, *severe chronic scarring*, consists of a finding **scarring** which is modified by a severity concept **severe** and a temporal concept **chronic**. d. **Related Finding**. A Related Finding is another finding which relates to the primary finding. This finding is nested in the main finding because of its relation with the main one. The definition of **Relational Finding** is not shown in this paper, but it consists of a slot whose value is a relation, such as **consistent with**, and a slot whose value is **Rad Finding Struct**, the structured form of the nested finding. f. The slots **Evidential Procedure**, **Technique Information**, and **Management Information** represent other types of information found in CXR's such as evidence of surgical procedures (**mastectomy**), information concerning technical issues related to the xray (**expiratory film**), and to management type of information (**followup recommended**).

Another important structure is the structure representing body locations associated with the findings. The concept Bodyloc is as follows:

```
[Bodyloc]-
  (Primary Loc)->[Bodyloc:{*}]
  (Spatial Mod)->[Spatial Relation:{*}]
  (Bodyloc Mod)->[Bodyloc:{*}]
  (Region Mod)->[Region:{*}]
  (Position Mod)->[Position:{*}]
  (Quantity Mod)->[Quantifier:{*}].
```

The meaning of the slots for Bodyloc are: a.**Primary Loc**. Represents the primary body location or region. In *heart is enlarged*, the primary loc is **heart**. b. **Spatial Mod**. Represents the prepositional or adverbial relation associated with the Primary Loc slot. If the radiology report states *opacity under left lung*, the spatial relation is **under**. c. **Bodyloc Mod**. Represents a body location modifier of the primary body location. In *finger of hand*, the primary location is **hand**, modified by a body location **finger**. d. **Region Mod**. Represents a region modifier of the primary body location. A region is a general area, such as **upper** or **mid**. e. **Position Mod**. Represents the orientation modifier of the primary body location. In *transverse heart*, the orientation modifier is **transverse**. f. **Quantity Mod**. Represents a quantifier, such as **2** in *2 fingers.*

The representation of the other concepts, such as **temporal**, **degree**, **certainty**, and **quantity** follow the same format, but are not shown in this paper. An example of the structured form for a finding in a CXR is shown in Section 5.

5 The Controlled Vocabulary

The MED forms a semantic network, where inheritance and multiple classification are standard. Every concept in the MED is a class, which has a certain position in the network. Following CG formalism, the hierarchical relations are listed separately from the definition of the concepts. Below we show some of the hierarchical information and some of the concepts in the MED. The ** indicates that the concept is lexically decompositional, which implies that all the words of the concept are not always contiguous to each other in the reports, and therefore the concept should be represented as a complex structure consisting of relations with other more basic concepts. Lexically simple concepts consist solely of a name which is not decomposable. Although we recognize that conceptual domain knowledge associated with medical concepts is critical to represent, we are omitting this level of knowledge from our discussion in order to concentrate on the issue of compositionality.

In describing the type hierarchy of the concepts the following form is used: Concept2 < Concept1. In the above form Concept2 is a subtype of Concept1. The root node of the network is **Medical Entity**.

```
Chest Xray Report < Medical Entity
Rad Finding Struct < Medical Entity
Procedure < Medical Entity
Rad Finding < Medical Entity
Descriptor < Medical Entity
Modifier < Medical Entity
```

```
Bodyloc < Modifier
cardiomegaly** < Rad Finding
enlarged < Descriptor
heart < Bodyloc
left lower lobe** < Bodyloc
Temporal < Modifier
increase < Temporal
```

The two lexically complex concepts that are shown are **cardiomegaly** and **left lower lobe**. Although cardiomegaly is one word, it is complex lexically because it consists of two morphemes, *cardio-* denoting *heart*, and *-megaly* denoting *enlarged*. In the reports, this concept may be expressed as *heart is enlarged* or other equivalent variant forms. Likewise, **left lower lobe** may occur in the reports, as in *left and right lower lobes*.

Below, we represent the compositional structure of the two complex concepts shown above. The compositional representation is crucial to our model, because it allows pieces of a concept (from a report or another controlled vocabulary) which are not necessarily contiguous to be composed to form the correct underlying concept.

```
[cardiomegaly]-
    (Central Finding)->[enlarged]
    (Bodyloc Mod)->[heart].

[left lower lobe]-
    (Primary Loc)->[lobe]
    (Region Mod)->[left]
    (Region Mod)->[lower].
```

Using this schema, it would be possible to map sentences such as *the heart appears enlarged* to the concept **cardiomegaly** even though the concept does not occur explicitly in the report. For example, the finding portion of the structured form for the above sentence would initially be as follows:

```
[Rad Finding Struct:#id]-
    (Central Finding)->[enlarged]
    (Bodyloc Mod)->[heart]
    (Finding Mod)->[appears]
```

Since the **Central Finding** and the **Bodyloc Mod** relations together are equivalent to the entire structured form of **cardiomegaly**, the value **enlarged** of the **Central Finding** will be changed to the concept **cardiomegaly**. Similarly, if the report consists of *infiltrate in left and right lower lobes*, the structure of the MED concepts **left lower lobe** and **right lower lobe** would be found as the corresponding MED concepts.

6 Discussion

It is important to stress the distinction between the concepts in the vocabulary and concepts used to represent information extracted from CXR's. A finding in a report is not the same as a generic concept in the MED, but rather is an instance of a concept. In general, a particular Chest Xray Report is an instance of the generic concept **Chest Xray Report**, and therefore is denoted as [Chest Xray Report:#], and a structured finding is an instance of the concept **Rad Finding Struct** and is therefore denoted as [Rad Finding Struct:#]; the # symbol represents that an identifier is associated with the structure. **Rad Finding Struct** consists of a **Central Finding** whose value is a concept which is in the MED, along with modifiers which are in the MED. If the value of **Central Finding** is part of a more complex concept in the MED, it must be mapped to that concept before being stored in the final encoded form, which in our system would be the coded relational patient database.

The ability to map from one controlled vocabulary to another is recognized as a critical task in Medical Informatics [4, 5, 11]. The components of a term could be structured manually but that would be very time-consuming. Because the structuring process is similar to that of mapping natural language to a controlled vocabulary, it is possible for us to perform this task automatically by using the same NLP methodology that structures the radiology reports. For example, the phrase *increased paramediastinal opacity* could be taken from another vocabulary and made a unique MED concept. It would be processed using the NLP processor to specify its compositional structure. Its structure would consist of the MED concept **opacity** with modifiers representing **increased** and **paramediastinum**. A variant expression which is equivalent, such as *marked increased opacity in the right paramediastinum* would be mapped to the MED concept **increased paramediastinal opacity** with modifiers because its structured form would be contained in the structure of that MED concept. An alternative approach may not incorporate the above complex concept into the MED. In that case the same concept could be retrieved from the structured forms generated from the reports by searching for a **Central Finding** with the value **opacity**, a **Finding Mod** with the value **increased**, and a **Bodyloc Mod** with the value **paramediastinum**.

The impression section of 8,000 chest x-ray reports obtained from the online clinical database at CPMC were automatically processed and mapped into the representational model presented in this paper. Generally, the model was complete enough to adequately represent the information in the reports. However, some types of information could not be represented completely. For example, in *bilateral, right greater than left, pleural effusions*, there was no way to represent the information that the effusion was greater on the right than on the left. However, the information that there were pleural effusions occurring on both sides could be represented. Thus, further work should be done in order to evaluate the model. In addition, the model should be extended to other domains to see how extensible it actually is.

The results presented here were derived from a combination of methods, some manual and some automated. The approach is one that combines a top-down view of analysis of the reports and a bottom-up view to identify the terms and their compositional structures. The resulting structures represent radiology findings in a way that is a) derivable through automated natural language

processing, b) consistent with the requirements of our Medical Entities Dictionary, c) consistent with the storage requirements of encoded data, and d) usable by automated decision support. Most importantly, the representational model consists of an accurate and verifiable structured form because it is not based on an ad hoc structural organization of the information but on the basic underlying semantic relations found in the domain. Because our approach is based on the semantics inherent in the domain itself, the same natural language text processing methodology that is used to process the reports is also used to represent the structure of lexically complex terms in the controlled vocabulary, and to translate terms from another controlled vocabulary into our controlled vocabulary.

Other groups have also independently developed conceptual models of CXR's as participants in the Harriman workshop [3] sponsored by the CANON group. One of the aims of CANON is to merge the different models into one standard model in order to facilitate the sharing of medical terminology and clinical data.

Acknowledgements

Dr. Friedman is supported in part by grant number R29 LM05397 from the National Library of Medicine and grant number 6-61483 from the Research Foundation of CUNY. Drs. Cimino and Johnson are supported in part by a Unified Medical Language System grant from the National Library of Medicine and by a development contract with IBM Corporation. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official view of the sponsors.

Appendix A

The following is a description of CG notation. In CGs a concept is similar to a frame, and a slot is similar to a relation. A concept is enclosed in square brackets, followed by the relations associated with it. Each relation appears in parentheses and is followed by an arrow (→). The slots are indented for readability. The values that each slot can take are specified by a *domain* concept that appears in square brackets after the arrow. Thus, the general format of a concept with N relations is:

```
[Concept]-
     (Slot1)->[Domain1]
     (Slot2)->[Domain2]
     :
     (SlotN)->[DomainN].
```

Note that the main concept is followed by a dash (-), and is terminated by a period (.). The number of values that a slot a permitted to have (its cardinality) is indicated by including a constraint C following the domain name. If C is :{*} the slot may have 0 or more values; if it is :@>1, the slot may have 1 or more values; if it is :@<2, the slot may have 0 or 1 values, and if it is :@1, the slot must have exactly 1 value.

References

[1] Archbold A and Evans D. On the topical nature of medical charts. In *Proceedings of the Thirteenth Annual Symposium on Computer Applications in Medical Care*, 1989.

[2] Mori A, Bernauer J, Pakarinen V, Rector A, et al. Models for representation of terminologies and coding systems in medicine. In *Proceeding of the Seminar: Opportunities for European and US Cooperation in Standardization in Health Care Informatics*, Geneva, September 1992.

[3] Evans D, Chute C, Cimino J, et al. Torwards a medical concept representation language for electronic medical records. In *Proceedings of the Third Annual Educational and Research Conference of the American Medical Informatics Assoc*, 1993. in press.

[4] Masarie F, Miller R, Bouhaddou O, Giuse N, and Warner H. An interlingua for electronic interchange of medical information. *Computers and Biomedical Research*, 24:379–400, 1991.

[5] Cimino J and Barnett G. Automated translation between medical terminologies using semantic definitions. *MD Computing*, 7(2):104–109, 1990.

[6] Cimino JJ, Hripcsak G, Johnson SB, and Clayton PD. Designing an introspective, multipurpose, controlled medical vocabulary. In *Proceedings of the 13th Annual Symposium on Computer Applications in Medical Care*, pages 513–518, Washington, D.C., 1989. IEEE Computer Society Press.

[7] Campbell K and Musen M. Respresentation of clinical data using SNOMED III and conceptual graphs. In *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care*, 1992.

[8] Huff SM and Warner HR. A comparison of Meta-1 and HELP terms: implications for clinical data. In R.A. Miller, editor, *Proceedings of the 14th Symposium of Computer Applications in Medical Care*, pages 161–165, 1990.

[9] J. F. Sowa. *Conceptual Structures.* Addison-Wesley, Reading, Mass, 1984.

[10] Nowlan W, Rector A, et al. A patient care workstation based on user centred design and a formal theory of medical terminology. In *Proceedings of the 15th Symposium of Computer Applications in Medical Care*, pages 855–857, 1991.

[11] Yang Y and Chute C. An application of least squares fit mapping to clinical classification. In *Proceedings of the Sixteenth Annual Symposium on Computer Applications in Medical Care*, pages 460–464, 1992.